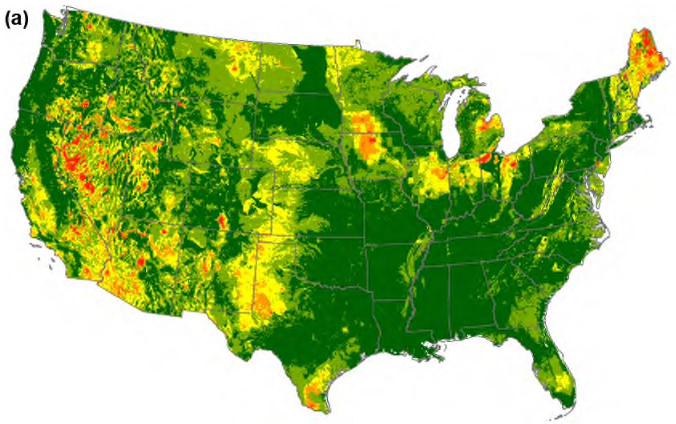


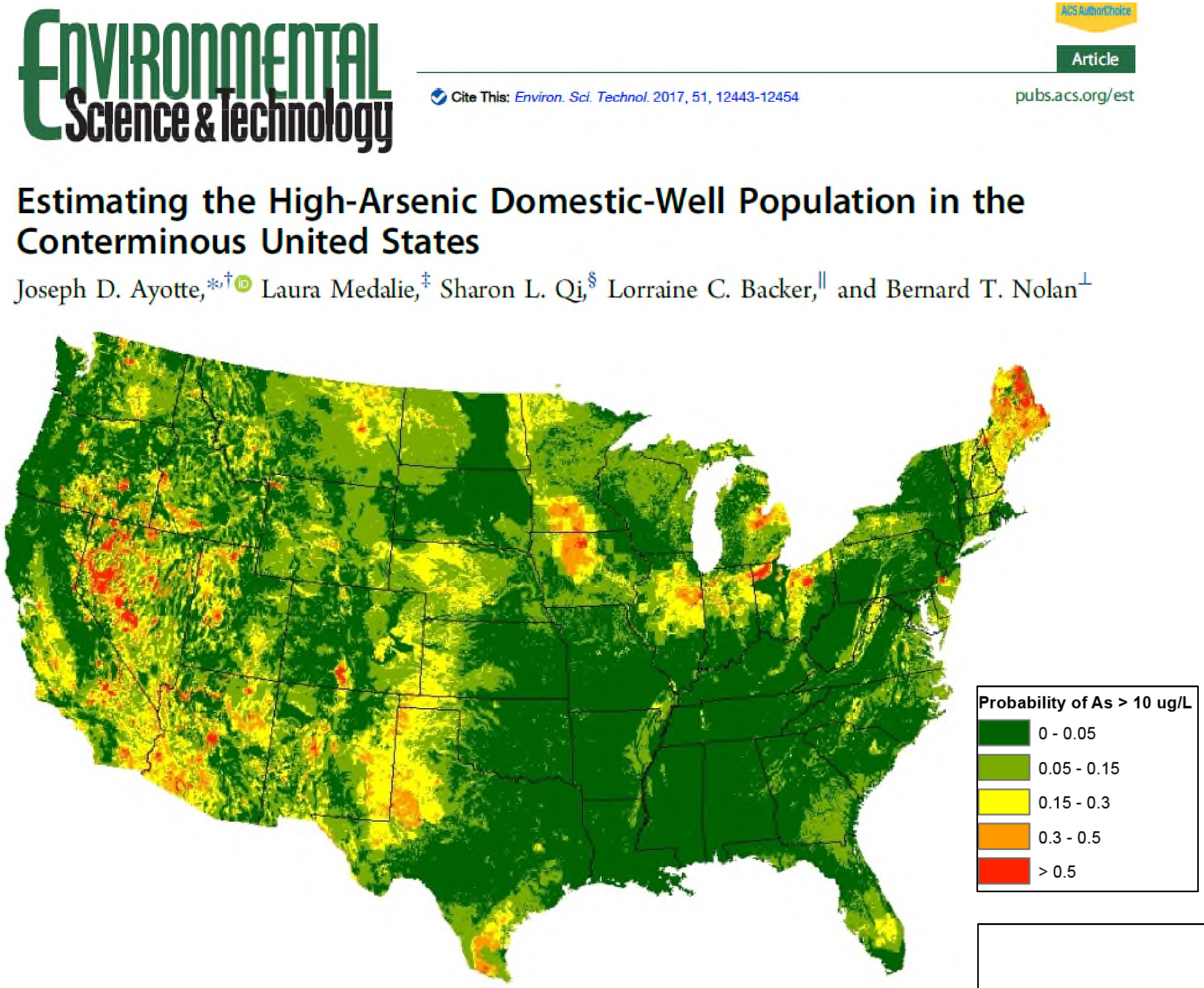
A comparison of statistical modeling techniques to predict arsenic in domestic wells in the CONUS

Melissa Lombard, Bernard T. Nolan, Mathew Gribble, Maria Argos, Joseph Ayotte

USGS science for a changing world

EMORY UNIVERSITY

THE UNIVERSITY OF ILLINOIS AT CHICAGO UIC



Environmental Science & Technology

ACS AuthorChoice

Article

Cite This: Environ. Sci. Technol. 2017, 51, 12443-12454

pubs.acs.org/est

Estimating the High-Arsenic Domestic-Well Population in the Conterminous United States

Joseph D. Ayotte,[*,†] Laura Medalie,[‡] Sharon L. Qi,[§] Lorraine C. Backer,[∥] and Bernard T. Nolan[⊥]

Probability of As > 10 ug/L

🟩	0 - 0.05
🟩	0.05 - 0.15
🟨	0.15 - 0.3
🟧	0.3 - 0.5
🟥	> 0.5

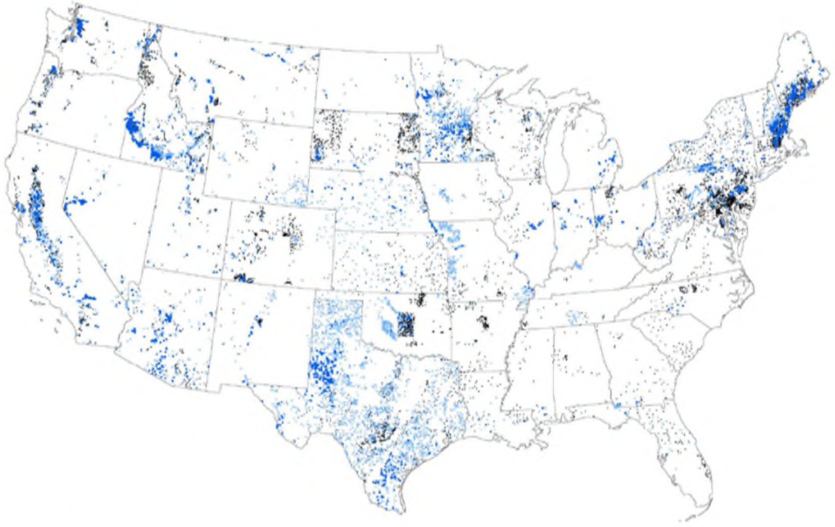
- Logistic Regression model
- Model response term is the probability of arsenic > 10μg/L

$$P(y|x) = \frac{e^{\beta_0+\beta_1 x_1 + \cdots + \beta_n x_n}}{1 + e^{\beta_0+\beta_1 x_1 + \cdots + \beta_n x_n}}$$

USGS
science for a changing world

Existing logistic regression model

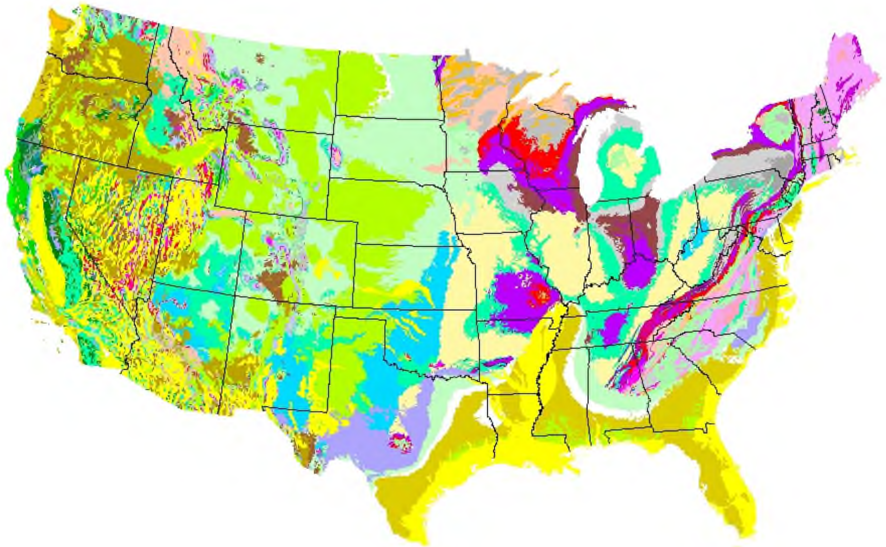
Arsenic concentrations
from 20,450 domestic wells



Black = As < 1 μg/L
Light blue = 1≥ As < 10 μg/L
Dark blue = As > 10 μg/L

From Ayotte et al. 2017, ES&T

42 predictor variables

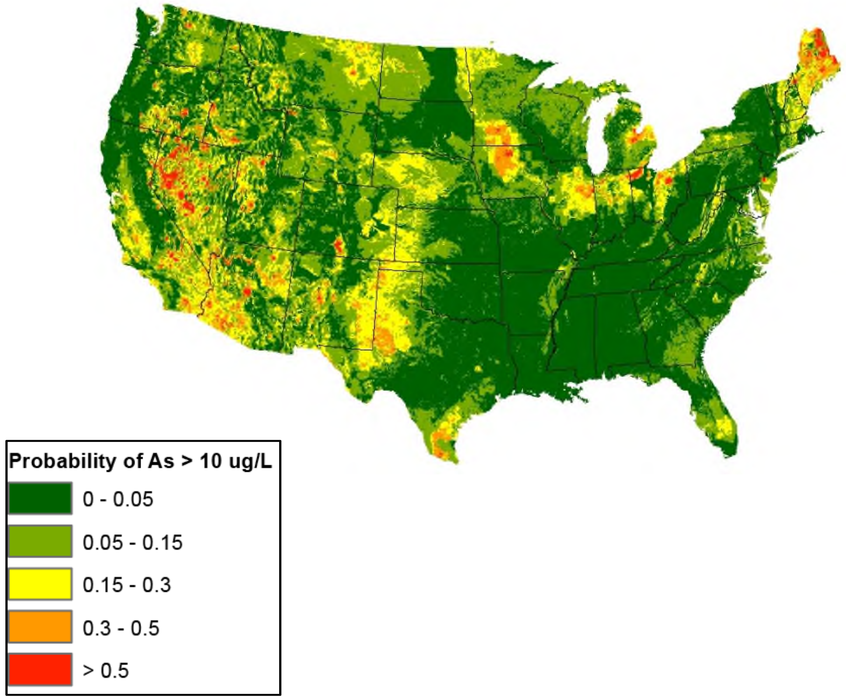


Generalized geology of the U.S.

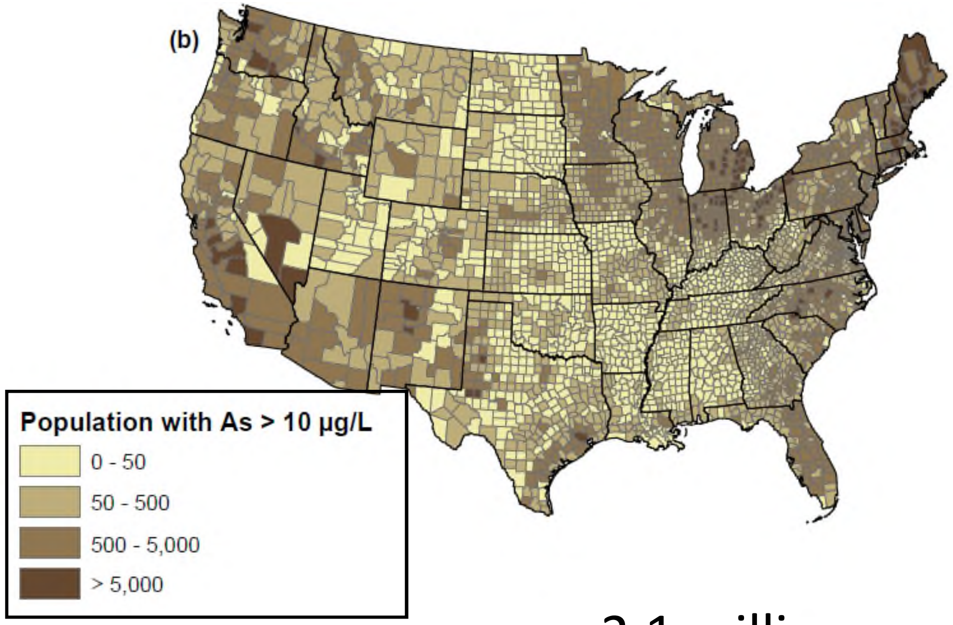
After King and Beikman, 1974

Existing arsenic model and exposure estimate

Arsenic model output



Arsenic exposure estimate



2.1 million people

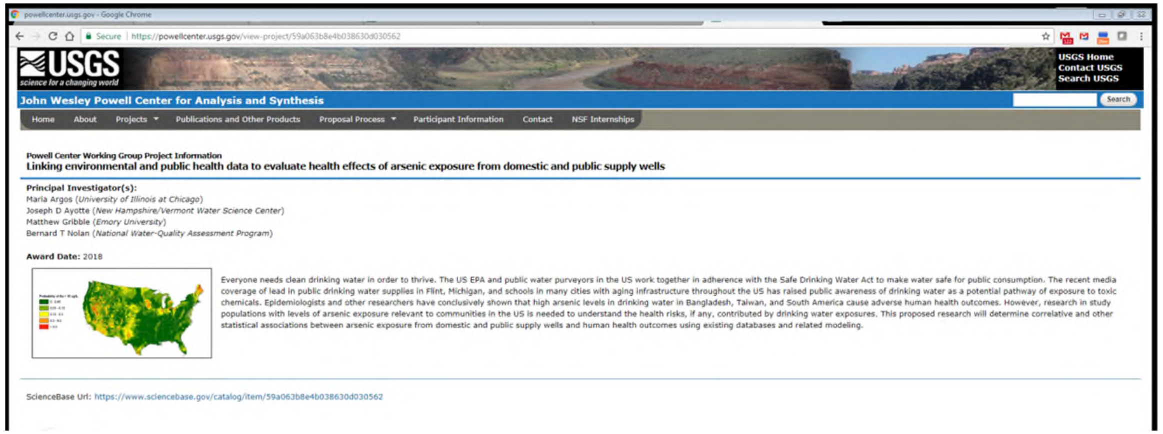
From Ayotte et al. 2017, ES&T

USGS Powell Center study

Linking environmental and public health data to evaluate health effects of arsenic exposure from domestic supply wells

- Make a new national arsenic model
  - Use machine learning methods
  - Update model variables

- Results will be used in epidemiology models to evaluate relationships between human health outcomes and arsenic in domestic wells.


powellcenter.usgs.gov - Google Chrome
Secure
https://powellcenter.usgs.gov/view-project/59a063b8e4b038630d030562
USGS
science for a changing world
USGS Home
Contact USGS
Search USGS
John Wesley Powell Center for Analysis and Synthesis
Search
Home
About
Projects
Publications and Other Products
Proposal Process
Participant Information
Contact
NSF Internships

Powell Center Working Group Project Information

Linking environmental and public health data to evaluate health effects of arsenic exposure from domestic and public supply wells

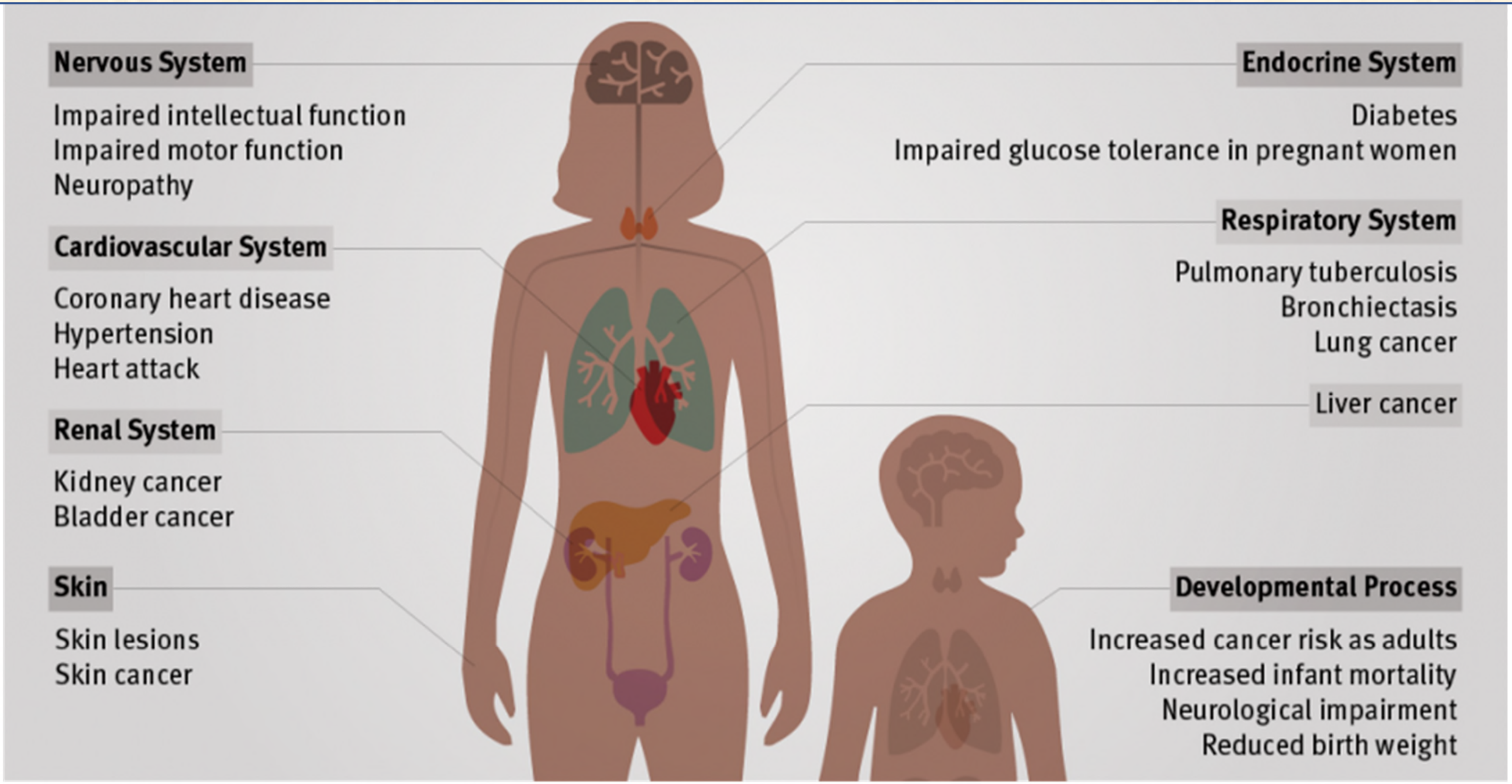
Principal Investigator(s):
Maria Argos (*University of Illinois at Chicago*)
Joseph D Ayotte (*New Hampshire/Vermont Water Science Center*)
Matthew Gribble (*Emory University*)
Bernard T Nolan (*National Water-Quality Assessment Program*)

Award Date: 2018

Everyone needs clean drinking water in order to thrive. The US EPA and public water purveyors in the US work together in adherence with the Safe Drinking Water Act to make water safe for public consumption. The recent media coverage of lead in public drinking water supplies in Flint, Michigan, and schools in many cities with aging infrastructure throughout the US has raised public awareness of drinking water as a potential pathway of exposure to toxic chemicals. Epidemiologists and other researchers have conclusively shown that high arsenic levels in drinking water in Bangladesh, Taiwan, and South America cause adverse human health outcomes. However, research in study populations with levels of arsenic exposure relevant to communities in the US is needed to understand the health risks, if any, contributed by drinking water exposures. This proposed research will determine correlative and other statistical associations between arsenic exposure from domestic and public supply wells and human health outcomes using existing databases and related modeling.

ScienceBase Url: https://www.sciencebase.gov/catalog/item/59a063b8e4b038630d030562

Arsenic's effects on the human body



Nervous System

Impaired intellectual function
Impaired motor function
Neuropathy

Cardiovascular System

Coronary heart disease
Hypertension
Heart attack

Renal System

Kidney cancer
Bladder cancer

Skin

Skin lesions
Skin cancer

Endocrine System

Diabetes
Impaired glucose tolerance in pregnant women

Respiratory System

Pulmonary tuberculosis
Bronchiectasis
Lung cancer

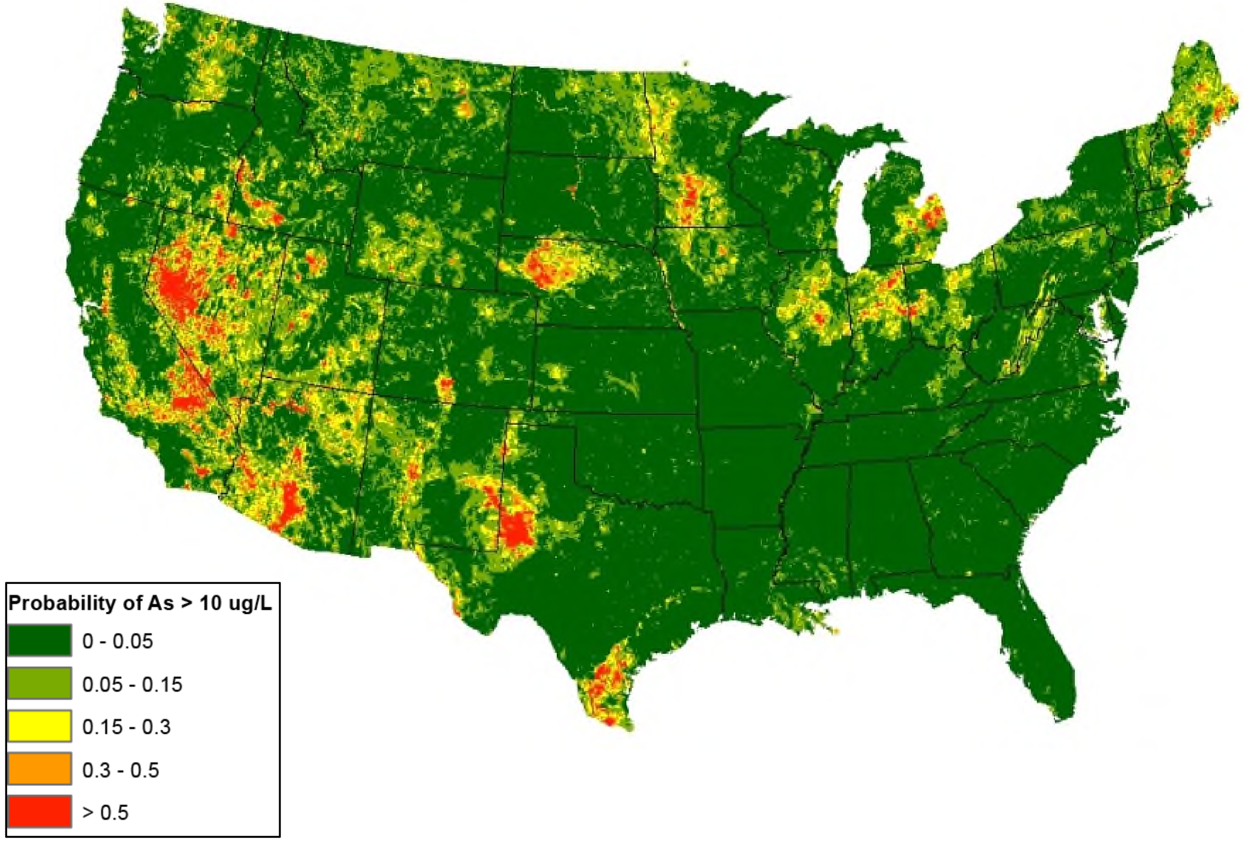
Liver cancer

Developmental Process

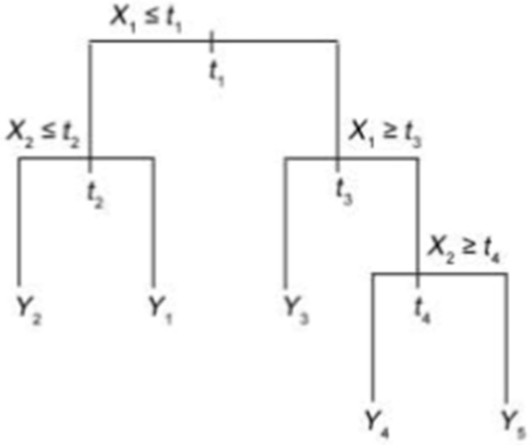
Increased cancer risk as adults
Increased infant mortality
Neurological impairment
Reduced birth weight

https://www.hrw.org/report/2016/04/06/nepotism-and-neglect/failing-response-arsenic-drinking-water-bangladeshs-rural#page

Boosted Regression Tree Model



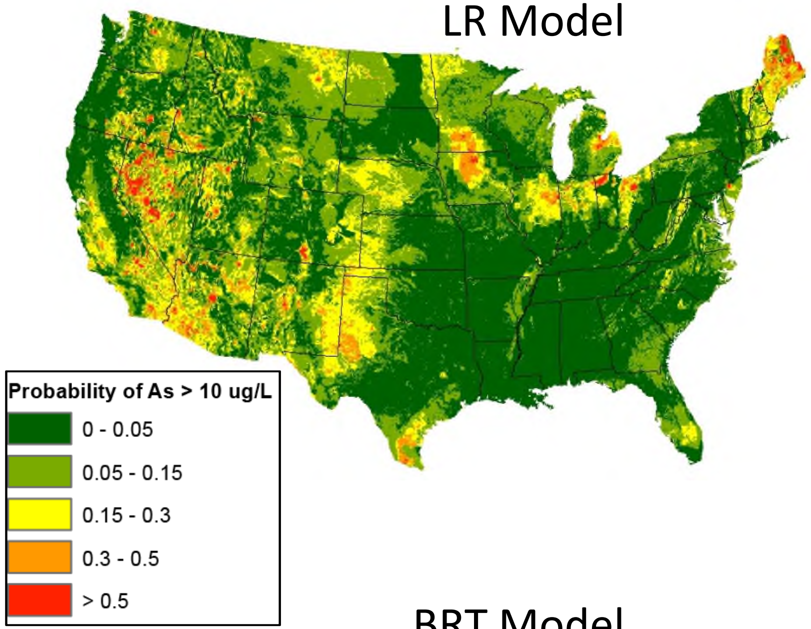
BRT Model
Number of trees = 4500
Interaction depth = 14
Learning rate = 0.01



From Elith et al., 2008

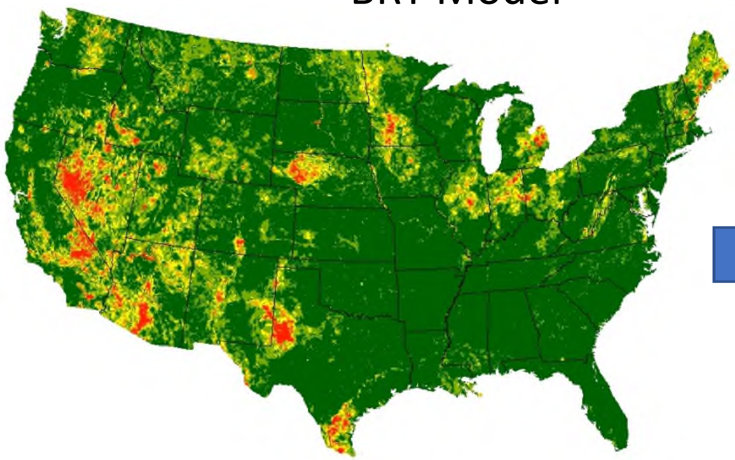
LR and BRT Model Comparison

LR Model

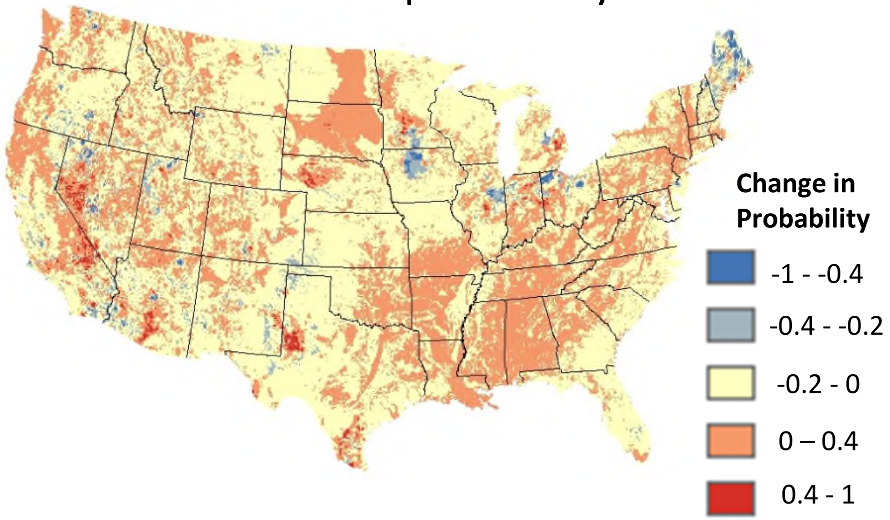


- Sharper delineation between high and low arsenic probabilities with BRT model

BRT Model



BRT minus LR probability



Model Comparison Predictive Performance

	Training Data			**Hold-out Data**		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
LR	89.9%	12.7%	99.3%	90.1%	13.9%	99.0%
BRT	95.6%	64.5%	99.3%	92.1%	43.7%	97.8%

Sensitivity = events (As>10μg/)

Specificity = non-events (As<10μg/L)

Random Forest Classification Model

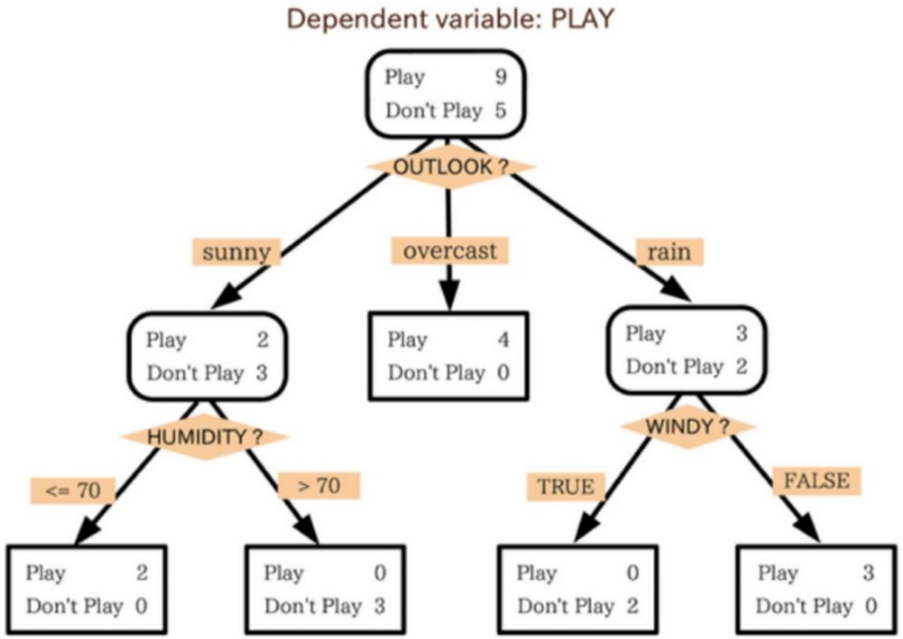
- Ensemble tree based machine learning method

- Model response term is a classification (category)
  - Arsenic ≤ 10 µg/L
  - Arsenic > 10 µg/L

RFC Model

Number of trees = 500

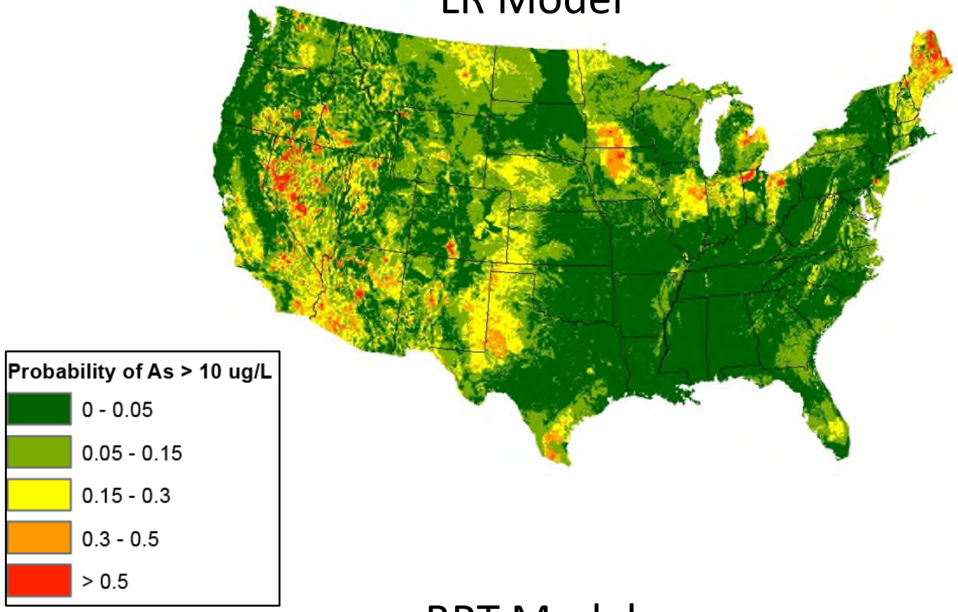
mtry = 38



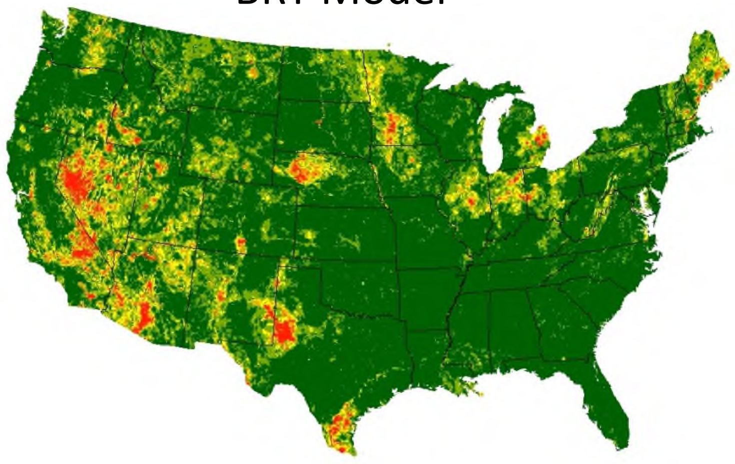
From medium.com

LR, BRT, & RFC Model Comparison

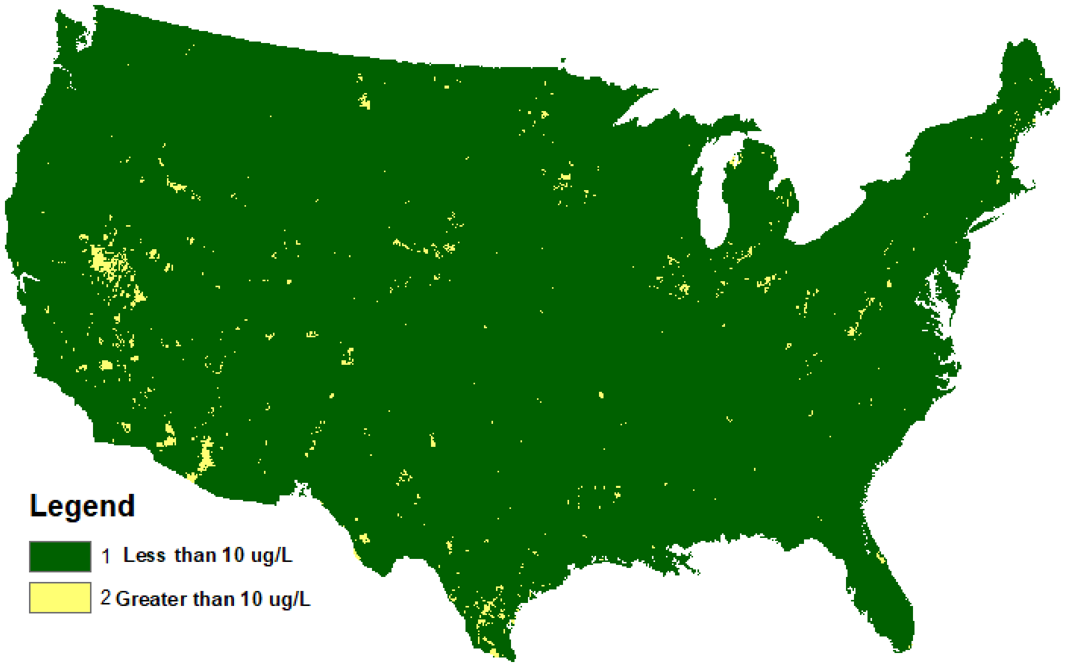
LR Model



BRT Model



RFC Model



Model Comparison Predictive Performance

	Training Data			Hold-out Data		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
LR	89.9%	12.7%	99.3%	90.1%	13.9%	99.0%
BRT	95.6%	64.5%	99.3%	92.1%	43.7%	97.8%
RFC 2C	99.9%	100%	99.3%	91.0%	38.4%	97.1%

Sensitivity = events (As>10μg/L)
Classification As>10μg/L

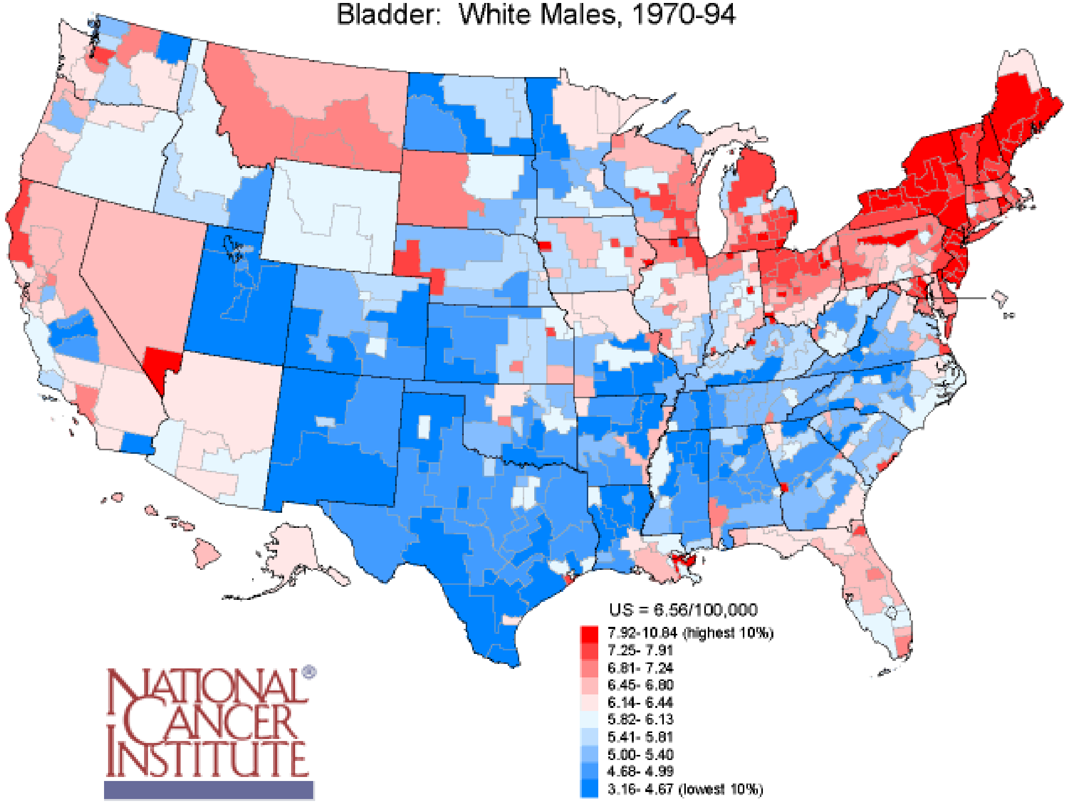
Specificity = non-events (As<10μg/L)
Classification As<10μg/L

Next steps

- Developing a RFC model with 4 concentration categories
  - ≤ 5 µg/L
  - 5 – 10 µg/L
  - 10 – 50 µg/L
  - > 50 µg/L

- Arsenic model results will be used in epidemiology models
  - Low birth weight
  - Cancers
  - Diabetes
  - Cardiovascular disease

Cancer Mortality Rates by State Economic Area (Age-adjusted 1970 US Population)
Bladder: White Males, 1970-94



NATIONAL CANCER INSTITUTE

CONTACT INFO:

Melissa Lombard, Ph.D.
US Geological Survey
331 Commerce Way
Pembroke, NH 03275
mlombard@usgs.gov
603-226-7816

Questions

?